Predicting Insurance Claim Filings

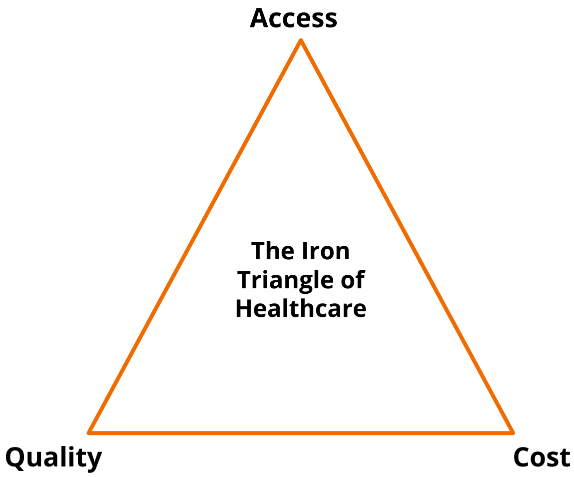
An Analysis of Demographic and Health Factors and Their Impact on Claim Size

By Stephanie Wairagu

DS-210

Motivation

- Health insurance is essential for any healthcare system
- Rising healthcare costs and chronic conditions affecting many individuals in the US have made managing insurance claims increasingly important for health insurance companies
- Analyzing claims data enables insurance companies to identify demand and utilization patterns and trends in medical care
- Claims data analysis also informs policy development



Objectives

This research aimed to:

1. Analyze the relationship between demographic and health factors and insurance claim filings

2. Predict the impact of these factors on the the size of claims filed

Data

Health insurance claims dataset containing demographic and health information about insurance policyholders and their claims

Each row represents a unique patient, claim observation

Each column represents a variable. Variables included:

> age, gender, BMI, blood pressure, diabetic status, smoking status, number of children, region & claim amount

Source: Kaggle

Exploratory Data Analysis

Examine variable data types

Check for missing or invalid data

Examine distribution of each variable

- Writing functions helped streamline the process

Rename yes/no values to:

- Smoker / Non-smoker
- Diabetic / Non-diabetic

Functions

Check for missing data

````
```{r missing-data}
missing_data <- function(data, col){
  data %>%
    summarize(n_missing = sum(is.na({{col}})))
}
```
````

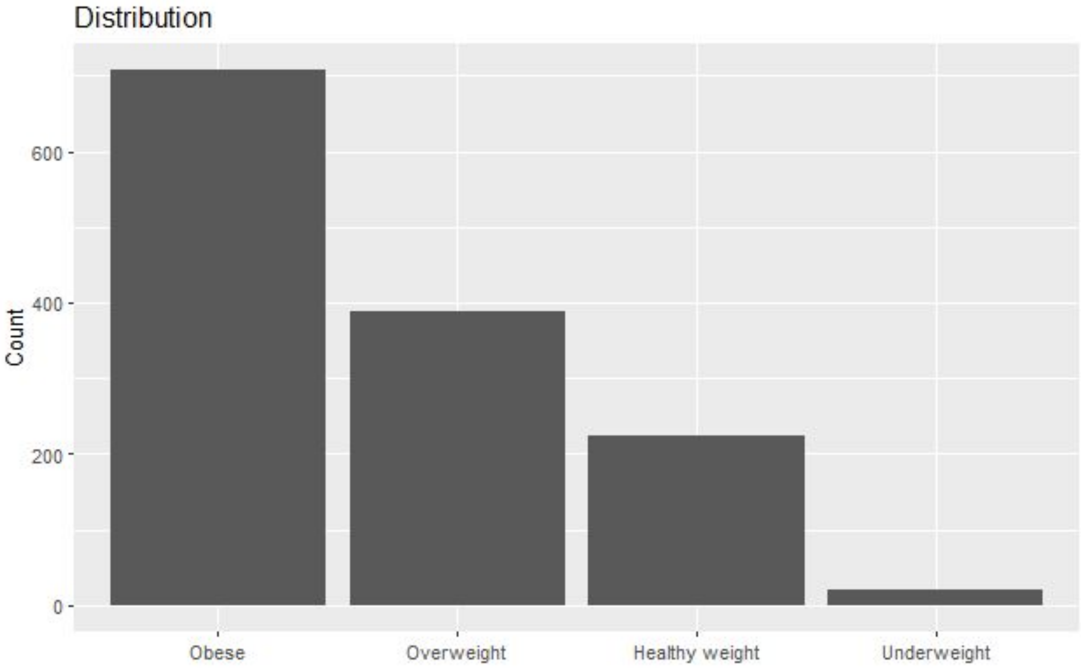
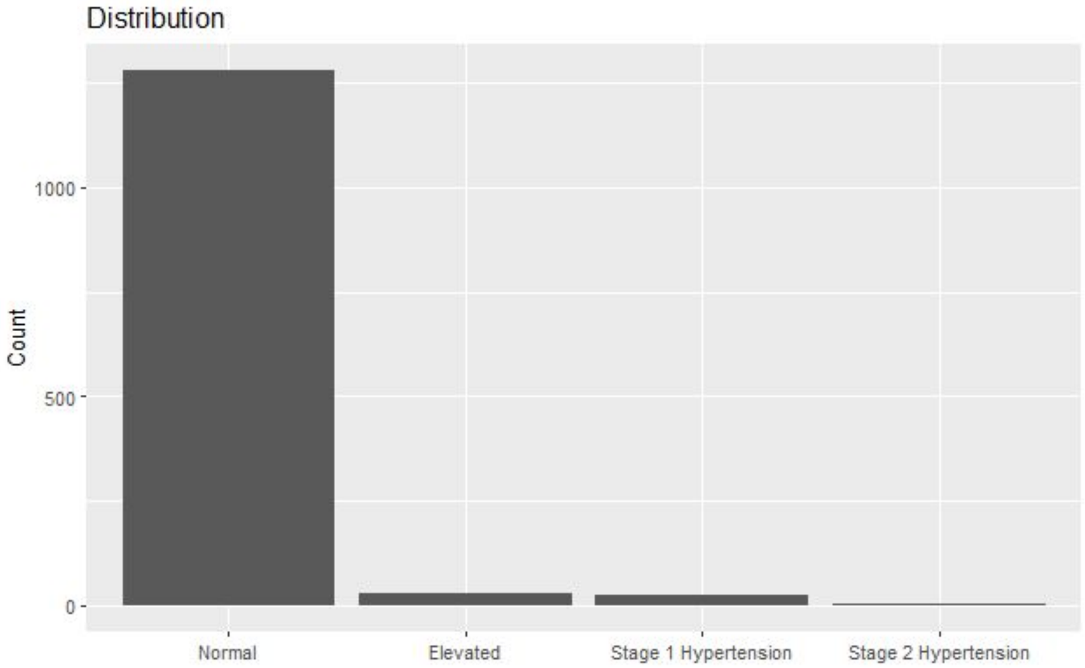
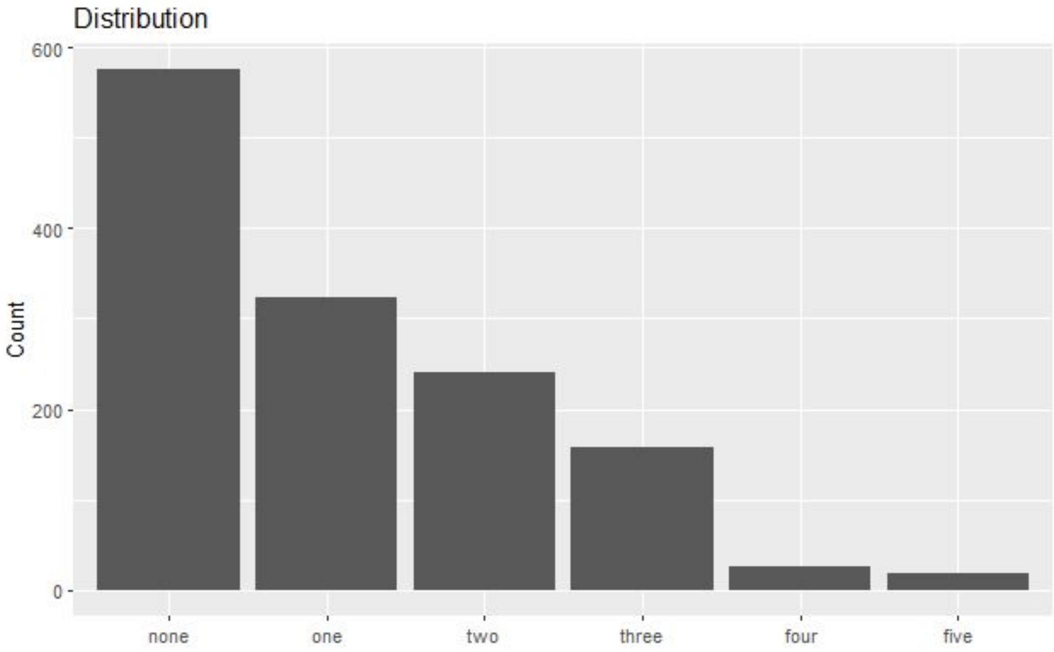
Functions

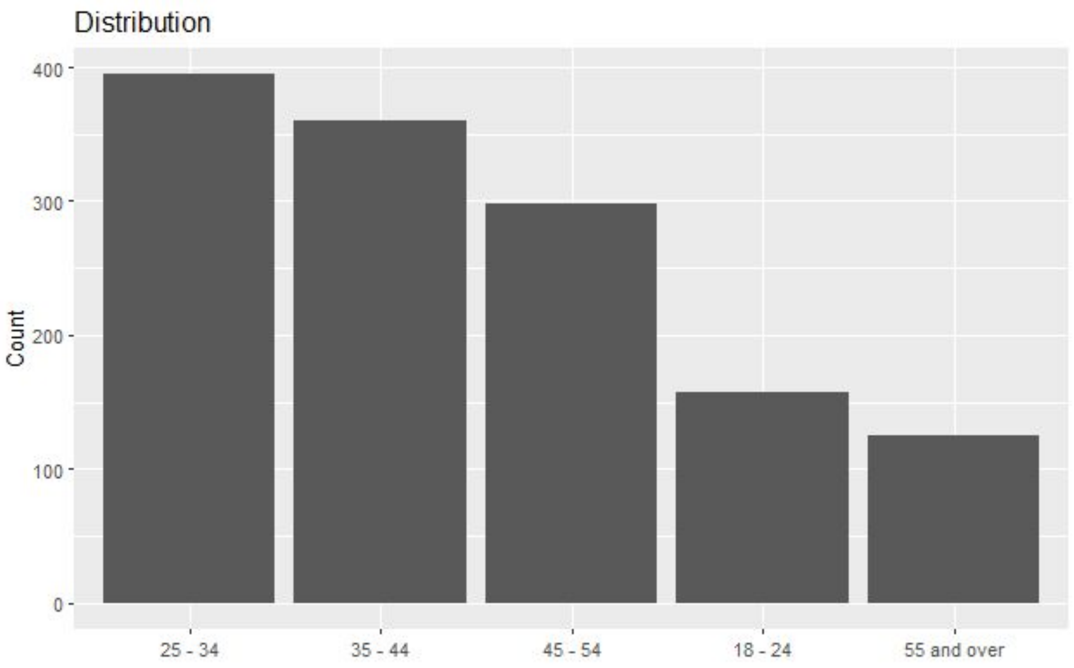
Check distribution of continuous variables

````
```{r dist-cont}
# Checking the distribution of continuous variables
hist_plot <- function(data, col){
  data %>%
    filter(!is.na({{col}})) %>%
    ggplot(aes(x = .data[[col]])) +
    geom_histogram(bins = 10) +
    labs(x = col, y = "Frequency", title = "Distribution")
}
```
````

Functions

Check distribution of categorical variables

```
```{r dist-cat}
# Checking the distribution of categorical variables
bar_plot <- function(data, col){
  data %>%
    filter(!is.na({{col}})) %>%
    count({{col}}) %>%
    ggplot(aes(x = reorder({{col}}, -n), y = n)) +          # -n to arrange bars in descending order
    geom_bar(stat = "identity") +
    labs(x = NULL, y = "Count", title = "Distribution")
}
```
```

New Variables

Weight categories

Blood pressure groups

Number of children - as a categorical variable

Age groups

Weight categories

```
```{r weight-category}
claims <- claims %>%
  mutate(weight_category = case_when(
    bmi < 18.5 ~ "Underweight",
    bmi < 25.0 ~ "Healthy weight",
    bmi < 30.0 ~ "Overweight",
    TRUE ~ "Obese"
  ))
```
```

```
```{r}
bar_plot(claims, weight_category)
```
```



A large number of claims are from patients who are obese, while very few are from underweight patients

Blood Pressure Groups

```
```{r bp-category}
claims <- claims %>%
  mutate(bp_category = case_when(
    bloodpressure < 120 ~ "Normal",
    bloodpressure < 130 ~ "Elevated",
    bloodpressure < 140 ~ "Stage 1 Hypertension",
    TRUE ~ "Stage 2 Hypertension"
  ))
```
```

```
```{r}
bar_plot(claims, bp_category)
```
```



A large majority of claims are from policyholders who have normal blood pressure, with a very small proportion being from those with elevated BP, stage 1 hypertension, or stage 2 hypertension

Number of Children

```
```{r children-cat}
claims <- claims %>%
  mutate(children_cat = case_when(
    children == 0 ~ "none",
    children == 1 ~ "one",
    children == 2 ~ "two",
    children == 3 ~ "three",
    children == 4 ~ "four",
    TRUE ~ "five"
  ))
```
```

```
```{r}
bar_plot(claims, children_cat)
```
```



The majority of claims are from policyholders who have no children, while the fewest are from those with five children

Age Groups

```
```{r age-cat}
claims <- claims %>%
  filter(!is.na(age)) %>%
  mutate(age_cat = case_when(
    age < 25 ~ "18 - 24",
    age < 35 ~ "25 - 34",
    age < 45 ~ "35 - 44",
    age < 55 ~ "45 - 54",
    TRUE ~ "55 and over"
  ))
```
```

```
```{r}
bar_plot(claims, age_cat)
```
```



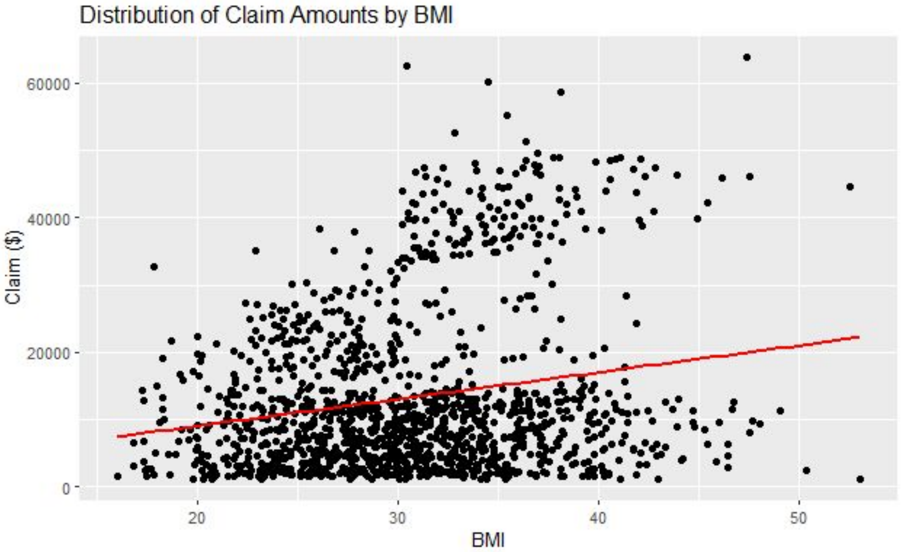
The majority of policyholders who filed a claim are between 25-34 years old
Data is more representative of young and early middle-aged adults

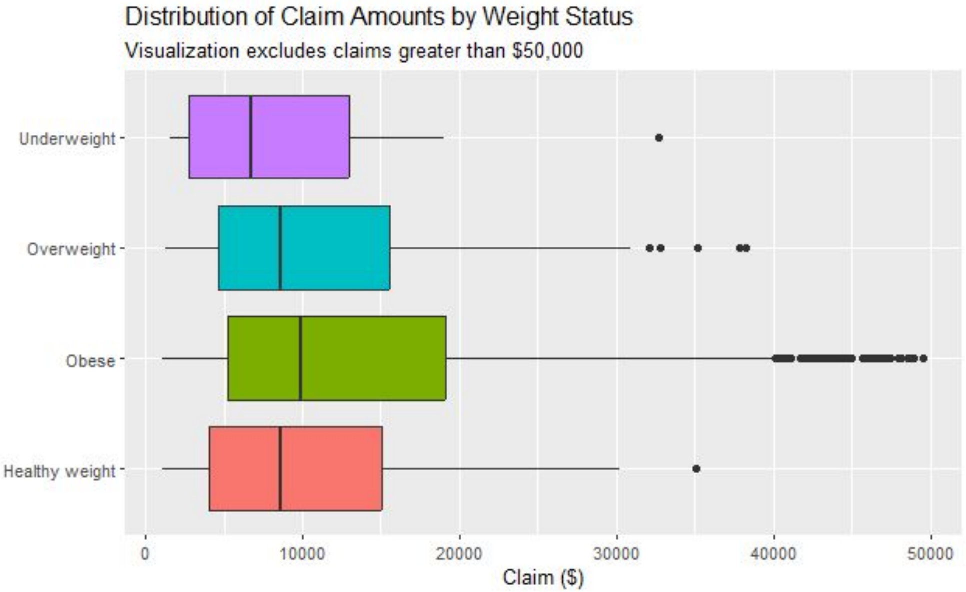
Bivariate Analysis

Exploring the relationship between:

1. Health factors and insurance claim amounts
2. Demographic factors and insurance claim amounts

Health Factors - BMI

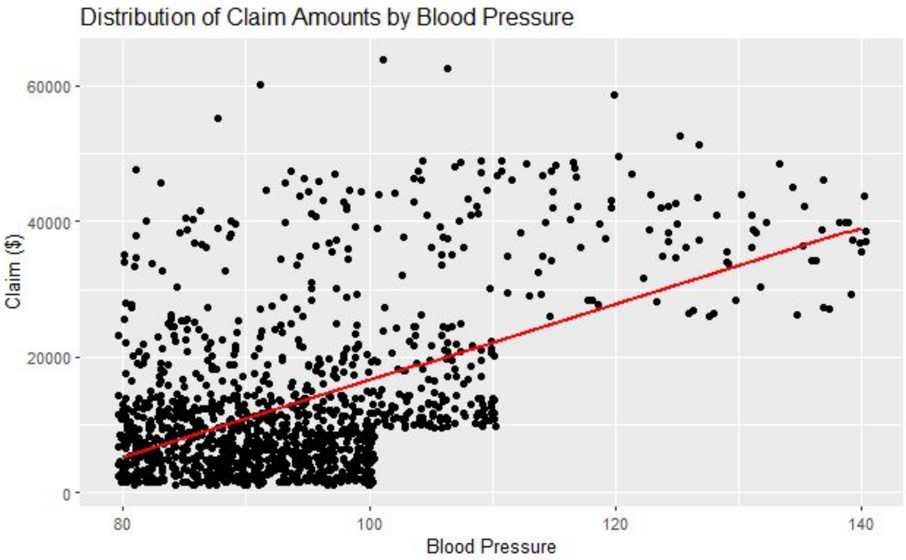


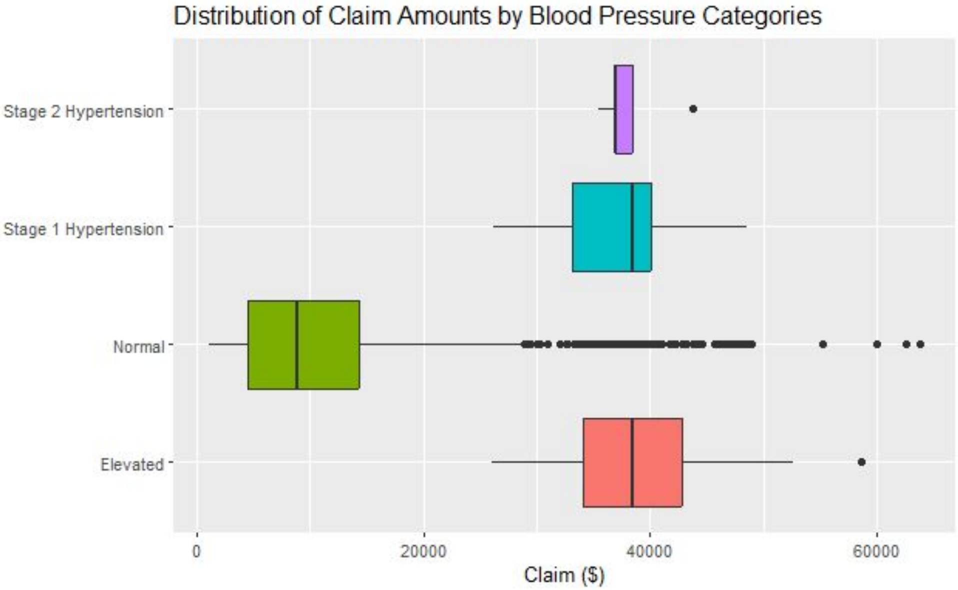


There is a positive linear association between BMI and claim amount

The underweight group has the lowest median claim amount while the obese group has the highest median claim amount

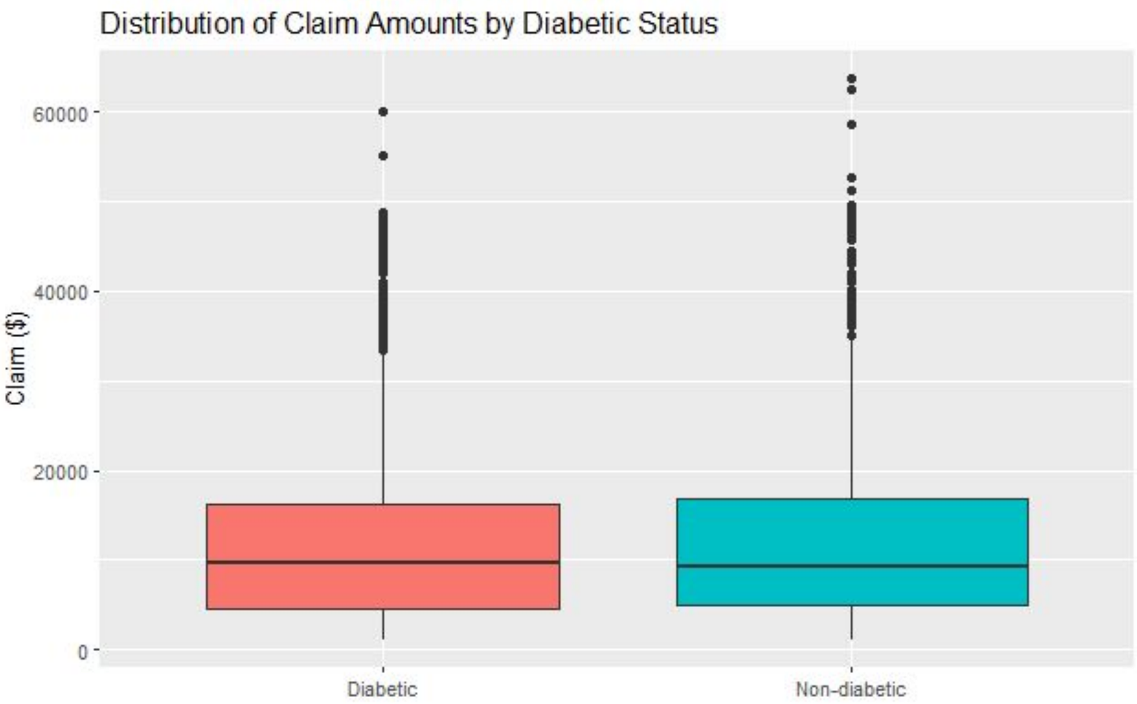
Health Factors - Blood Pressure





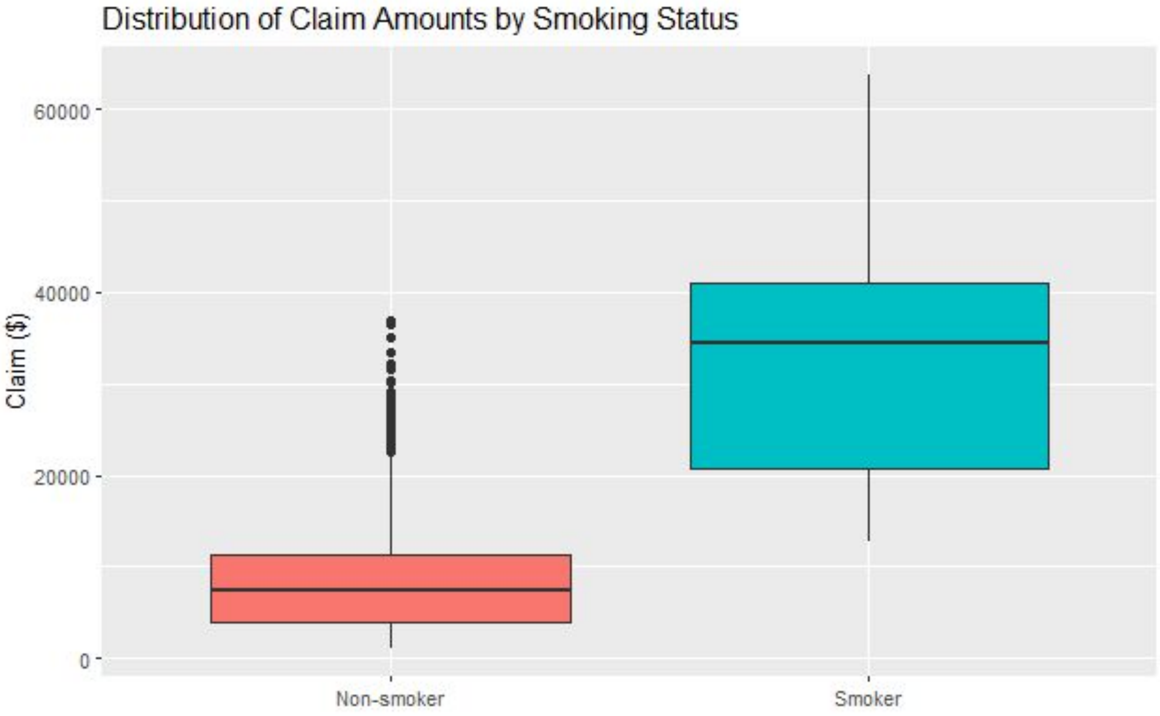
There is a positive linear association between blood pressure and claim amount
The median claim amount for those with normal blood pressure is much lower compared to the other categories

Health Factors - Diabetic status



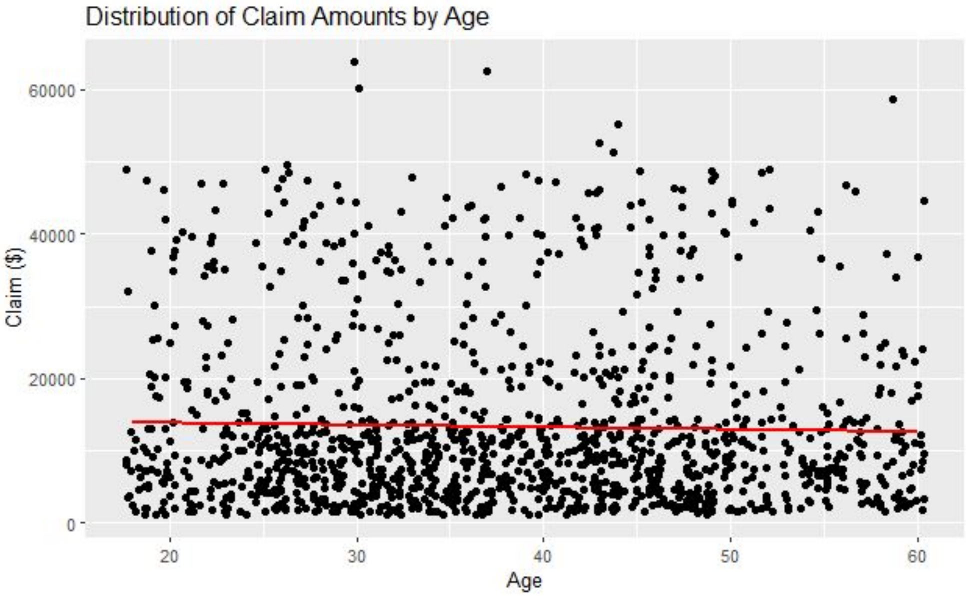
The distribution of claims for both the diabetic and non-diabetic policyholders appears to be fairly balanced
The median claim amounts for both groups seems to be similar

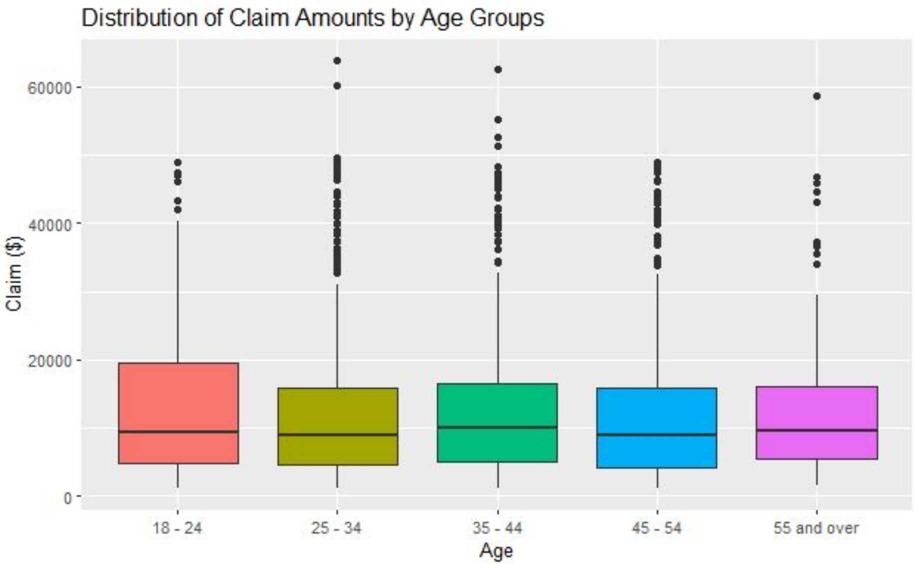
Health Factors - Smoking status



The distribution of claims is left-skewed for policyholders who are smokers
The median claim amount is much lower among non-smokers
Claim amounts for some of the non-smoker outliers are still lower than the median claim amount for the smoker group

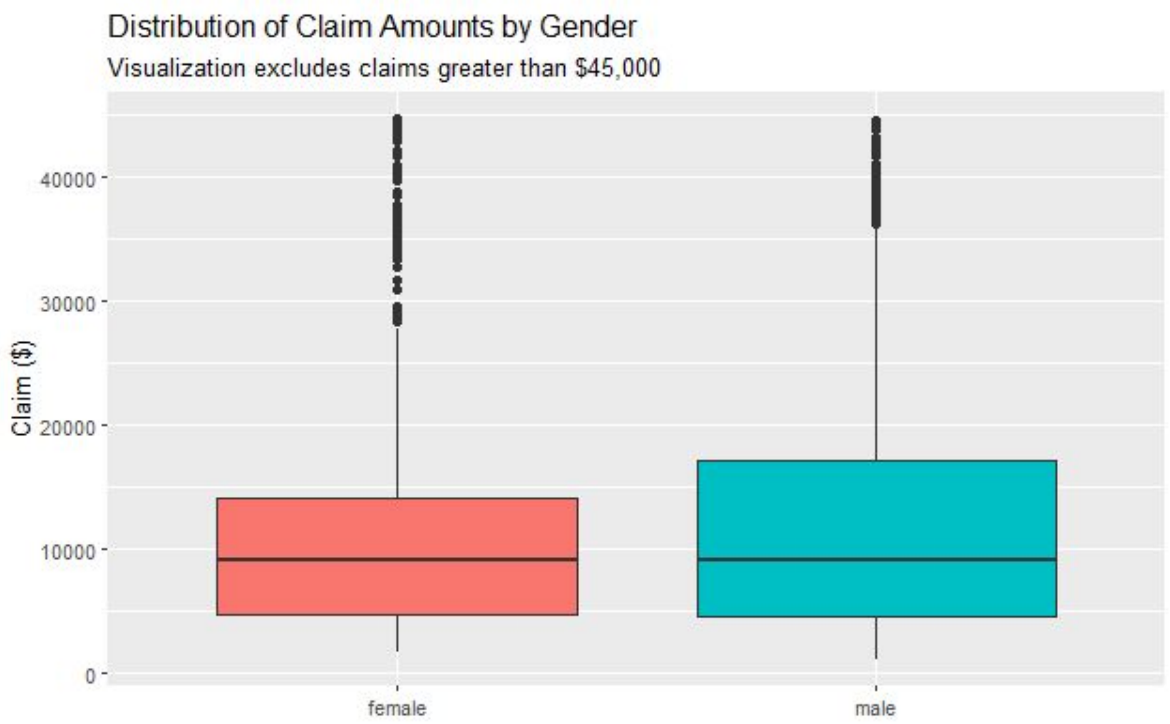
Demographic Factors - Age





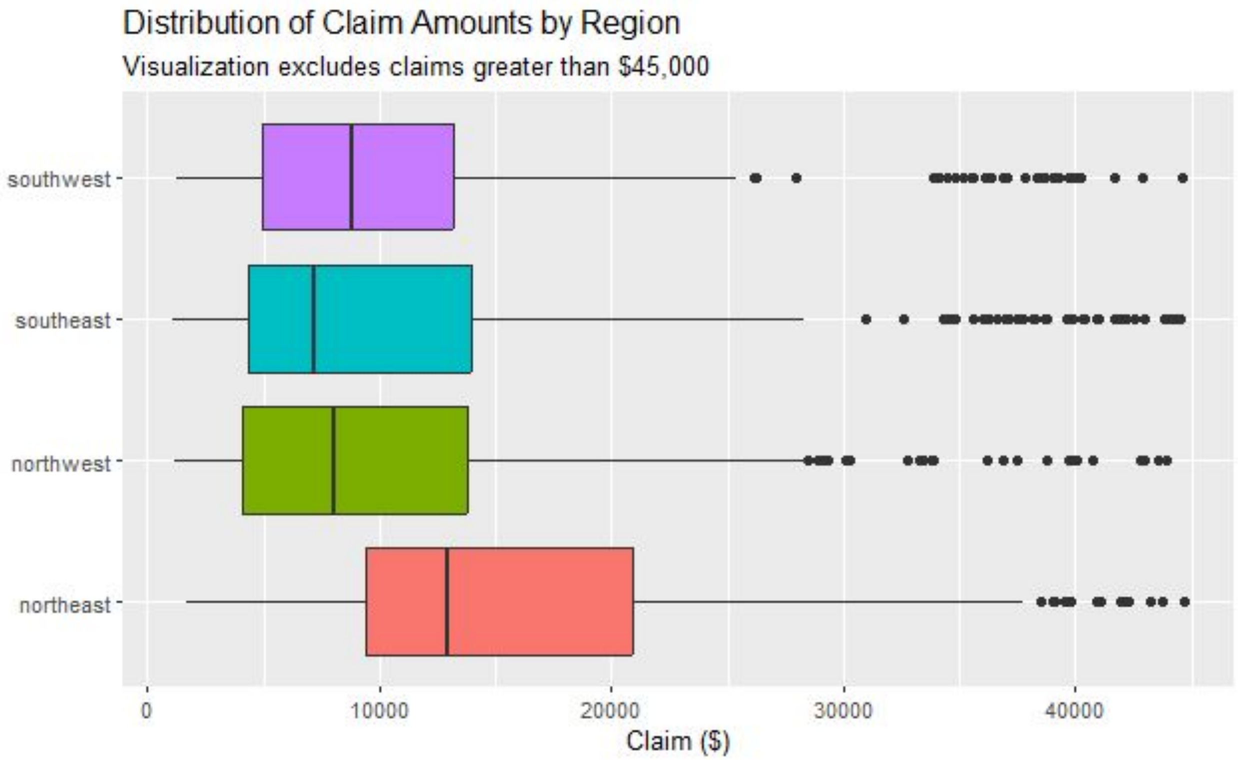
There is no significant relationship between age and claim amount
The median claim amounts appear fairly similar among the age groups

Demographic Factors - Gender



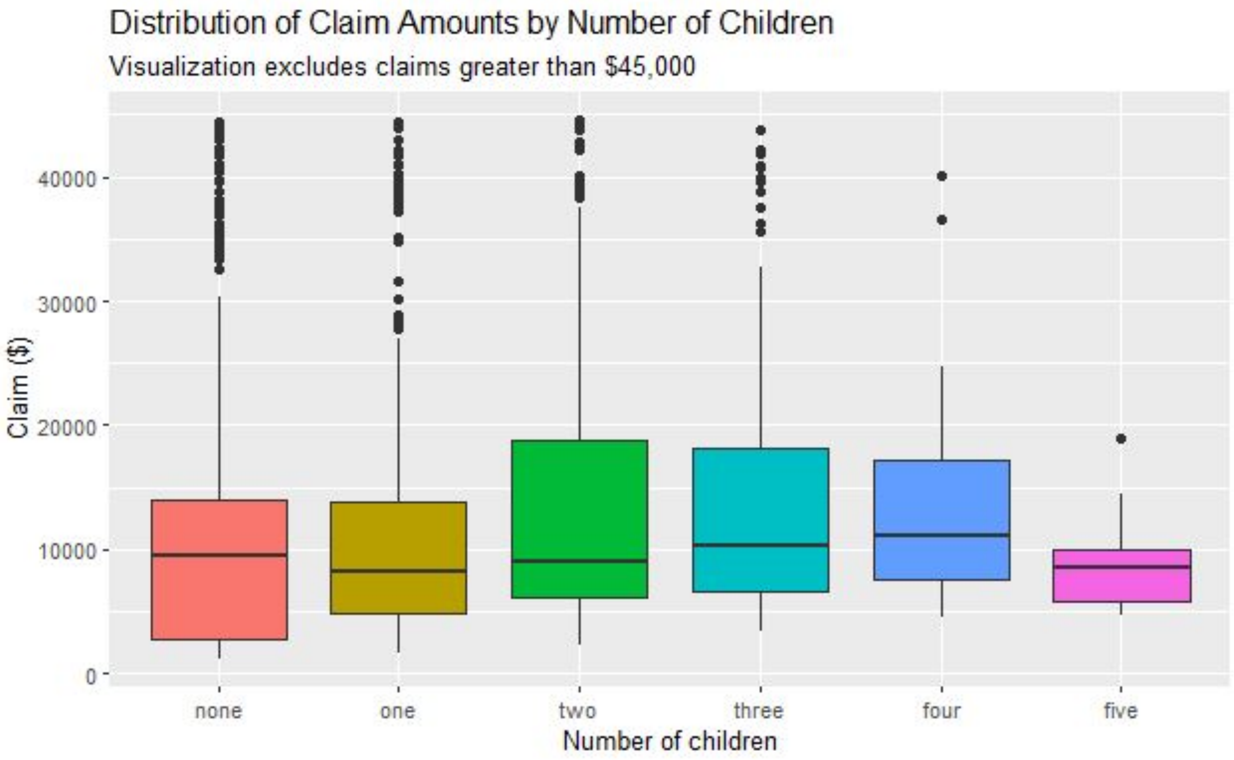
Median claim amounts appear to be very similar for both genders

Demographic Factors - Region



The northeast region has the highest median claim amount while the southeast region has the lowest median claim

Demographic Factors - No. of Children



The distribution of claims is left-skewed for policyholders with five children and right-skewed for policyholders with no children and those with two, three, or four children

The median claim amount is slightly higher for those with four children

Regression Analysis

Interested in predicting the relationship between the various health and demographic variables and claim amounts

From the initial EDA, we saw:

- There is an association between BMI and claim amount
- There is a strong association between blood pressure and claim amount
- There is no significant association between diabetic status and claim amount
- There is a strong association between smoking status and claim amount
- There is no significant association between age and claim amount
- There is no significant association between gender and claim amount
- There is an association between region and claim amount
- There might be an association between number of children and claim amount

Modeling

Test for collinearity (GGally)

Started with a full model

- Age, gender, and diabetic status were not significant
- p-values > 0.05
- Confidence intervals contained 0

Then used backward elimination to achieve my final model

Final Model

$$\widehat{claim} = -22020.82 + 228.78 bloodpressure + 351.85 bmi + 20649.65 smoker_{smoker} + 677.54 children - 1943.35 region_{northwest} \\ - 2880.72 region_{southeast} - 2221.61 region_{southwest}$$

Adjusted R^2 is 0.7049; hence, 70.49% of the variation in claim amounts can be explained by the model

The intercept is -22020.82. This represents the expected insurance claim amount for a patient who has a blood pressure of zero, a BMI of zero, is not a smoker, has zero children, and lives in a region that is not included in the model:

- Since this hypothetical patient is not realistic, the intercept value is not practical.

Conclusions

- Holding all other variables constant, the expected insurance claim amount increases by $228.78 for every unit increase in blood pressure.
- Holding all other variables constant, the expected insurance claim amount increases by $351.85 for every unit increase in BMI.
- Holding all other variables constant, the expected insurance claim amount for a patient who smokes is $20,649.65 higher than for a patient who doesn't smoke.
- Holding all other variables constant, the expected insurance claim amount increases by $677.54 for each additional child the patient has.

$$\widehat{claim} = -22020.82 + 228.78 bloodpressure + 351.85 bmi + 20649.65 smoker_{smoker} + 677.54 children - 1943.35 region_{northwest} - 2880.72 region_{southeast} - 2221.61 region_{southwest}$$

Conclusions

- Holding all other variables constant, the expected insurance claim amount for a patient in the Northwest region is $1943.35 lower than for a patient in the Northeast region.
- Holding all other variables constant, the expected insurance claim amount for a patient in the Southeast region is $2880.72 lower than for a patient in the Northeast region.
- Holding all other variables constant, the expected insurance claim amount for a patient in the Southwest region is $2221.61 lower than for a patient in the Northeast region.
- It also appears that smoking status is the most important predictor variable in the regression model. It has the largest coefficient estimate and lowest p-value. (If removed from the model, adjusted R^2 drops to 0.3152)

$$\widehat{claim} = -22020.82 + 228.78 bloodpressure + 351.85 bmi + 20649.65 smoker_{smoker} + 677.54 children - 1943.35 region_{northwest} - 2880.72 region_{southeast} - 2221.61 region_{southwest}$$

Future Research

1. Since smoking status is the most important predictor of claim amounts, I'm curious about what aspect of smoking has a bigger impact on healthcare costs

Next steps would be to explore public health data sources that have information on smoking behavior and healthcare costs

- National Health Interview Survey
- National Health and Nutrition Examination Survey
- Behavioral Risk Factor Surveillance System

2. Conduct further analysis to determine whether there are interactions between predictor variables (e.g., between smoking status and age) that may influence insurance claim amounts